DOCUMENT RESUME

ED 212 673 TM 820 120

AUTHOR Moy, Raymond H.
TITLE Proficiency Standards and Cut-Scores for Language Proficiency Tests.
PUB DATE Mar 81
NOTE 30p.; Paper presented at the Annual Meeting of TESOL (Detroit, MI, March 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cutting Scores; Higher Education; *Language Proficiency; Norm Referenced Tests; *Scoring Formulas; *Standards; Testing Problems

ABSTRACT

The problem of standard setting on language proficiency tests is often approached by the use of norms derived from the group being tested, a process commonly known as "grading on the curve." One particular problem with this ad hoc method of standard setting is that it will usually result in a fluctuating standard dependent on the particular group being tested. In this paper, four main steps are presented in the setting of a non-arbitrary cut-score. They are: (1) checking to see that the test covers the abilities of the test population; (2) establishing a proficiency standard; (3) determining the best cut-score, given that there will be errors of placement; and (4) using external criteria for checking the proposed standards. Methods for carrying out each step are presented along with illustrative examples based on data from the UCLA English as a Second Language Placement Examination. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED212673

Proficiency Standards and Cut-Scores for Language Proficiency Tests

by

Raymond H. Hoy

University of California, Los Angeles

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Hoy

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of TESOL, Detroit, March 1981

TM 820 120

Abstract

The problem of standard setting on language proficiency tests is often approached by the use of norms derived from the group being tested, a process commonly known as "grading on the curve." One particular problem with this ad hoc method of standard setting is that it will usually result in a fluctuating standard dependent on the particular group being tested. In this paper, four main steps are presented in the setting of a non-arbitrary cut-score. They are: 1) checking to see that the test covers the abilities of the test population, 2) establishing a proficiency standard, 3) determining the best cut-score, given that there will be errors of placement, and 4) using external criteria for checking the proposed standards. Methods for carrying out each step are presented along with illustrative examples based on data from the UCLA English as a Second Language Placement Examination.

Much of the research in language proficiency testing has concentrated on questions related to constructing valid measures of proficiency. For example, Are integrative tests better than discrete point tests? or Do deletion rates and scoring methods affect the reliability and validity of cloze scores? More often than not validity is assessed via correlational evidence, but none of this evidence seems to shed any light on the practical problem of how the test results should be used. If any decisions are to be made, such as placement in English courses, admission to a university, or the awarding of scholarships to study abroad, a proficiency standard and a cut-score must be arrived at. It is probably in this search for proficiency standards that the ultimate validity of a test will be shown. If the test administrator uses a non-arbitrary method of setting a cut-score, it will force him to carefully consider the content of his test and specify how student performance on test items relates to the decision to be made. This is not an easy task and the task's difficulty is probably the reason for the most popular method of standard setting - the use of norms. That is, giving passing status to some upper percentage of test scores. This is the time-worn procedure of grading on the curve which, if used indiscriminately, will result in a fluctuating proficiency standard that depends on the group of persons taking the test, and more subtly on the particular test used for each test session.

The overall goal of following a particular proficiency standard setting procedure is to arrive at a cut-score that is acceptable to the test-takers and defensible by the test administrators. Cut-scores which are arbitrarily set will sometimes fail in the first regard and always fail in the second.

The purpose of this presentation is to go over the steps one should follow in setting a cut-score, to present some methods for carrying out those steps, and to provide examples wherever possible.[1] In general, there are four main steps one should go through: 1) checking to see that the test covers the abilities of the test population, 2) establishing a proficiency standard, 3) determining the best cut-score, given that there will be errors of placement, and 4) using external criteria for checking the proposed standards. None of the models to be presented will make the task of proficiency standard setting easy, but, hopefully, they will make the process less arbitrary.

To provide some context for these steps, consider the norm-referenced model just mentioned. In order to justify its use, two conditions must be satisfied. First of all one must assume the test population remains the same from testing session to testing session, and, second, the test

---

[1] The author would like to thank Earl Rand and the ESL section of the UCLA English Department for granting access to their data banks upon which the examples used in this paper are based.

forms must be parallel, that is, there should be no significant test by person interaction whereby test forms order persons differently beyond what could be expected due to measurement error fluctuation. How well these conditions are fulfilled can only be inferred by logical means and by inspecting the empirical data.

Checking the Appropriateness of the Test

The first step, then, in setting a norm-referenced cut-score is to look at the distribution of scores, and check to see if it is consistent with how one would expect ability to be distributed in the population. For example, Figures 1-3 illustrate three distributions of scores derived from sets of items in the FALL 1977 UCLA English as a Second Language Placement Examination, given to all entering foreign students. One distribution is from a 50 item cloze test. The next distribution is from a 50 word dictation test, and the third distribution is from 88 multiple-choice items covering reading, listening, and structure subtests. The total number of scores in each of the distributions is 460. With the cloze test it appears that there is a floor effect, but the spread of scores at the upper end seems okay. On the dictation, however, the distribution is quite different with a distinct ceiling effect. The test doesn't seem to discriminate among the high scorers. Also, on the other end of the scale are several persons with scores of

zero. In the last distribution, there is a negative skew, but it appears that the test can spread scores out on both ends of the scale. It is this distribution that seems to be the most consistent with how one would expect language ability to be distributed in a population where non-random selection is occurring. That is, one would expect that foreign students entering an American university would more than likely be selected in part on the basis of their English language ability thus resulting in the negative skew.

INSERT FIGURES 1-3 ABOUT HERE

To give some idea of the stability of these distributions, Figures 4-6 show the spread of scores from the corresponding subtests in the Fall 1978 testing sessions. At this time there were 689 examinees. The dictation and cloze passages were changed completely, and the multiple-choice test remained essentially the same with the addition of some new structure items. Even though the cloze passage was shortened to 25 items, its score distribution resembles that of the previous year with a normal-type distribution at the upper end of the scale and a floor-effect on the other end. The dictation distribution

is perhaps the most interesting in that it exaggerates the tendencies of the 1977 dictation test to have both a floor and a ceiling effect. It appears that for this test a student does either very well or very poorly. Because of the great difference in distributions from the 1977 to the 1978 dictations, a norm-referenced cut-score in this case probably does not have a stable interpretation. For the multiple-choice and cloze tests, the distributions strongly resemble those of 1977 and are consistent with the assumptions of a stable test population as well as parallel test forms. In Tables 1 and 2 an examination of the intercorrelations among the three tests reveals a high relationship between the multiple-choice and the dictation in 1977 and a moderately high and consistent relationship between multiple-choice and cloze in both years.

INSERT FIGURES 4-6 ABOUT HERE

INSERT TABLES 1 & 2 ABOUT HERE

Establishing the Proficiency Standard

Given that the test administrator is fairly satisfied that the observed distribution of scores is a function of how language ability is distributed in the population, the relatively difficult step of selecting a proficiency standard is reached. It should be noted that a distinction is made between a proficiency standard and a cut-score. While, a proficiency standard is an idealized level of ability which represents the proportion of the skill domain that the proficient examinee has mastery over, cut-scores are set in accordance to the particular test that is being used. So, although the proficiency standard may remain the same, a cut-score will vary with such things as the difficulty of the test, the reliability of the scores, and the different risks the administrator is willing to take in making decision errors - for example, one might feel that passing a non-proficient student is a far greater error than failing a proficient one.

Establishing a proficiency standard requires expert judgement and empirical verification. One particular methodology, suggested by Nedelsky (1954), is applicable to multiple-choice tests. Several raters independently review each item and determine which distractors a minimally proficient examinee is able to reject. The reciprocal of the remaining choices would then equal the chance of this student getting this item correct. For example, here are

two items where the best answer must be selected to complete the sentences:

1) _____ are they testing in the examination?
   a) Where
   b) Which
   c) When
   d) What

2) They have lived in Detroit _____.
   a) twenty years ago
   b) for twenty years
   c) during twenty years
   d) in twenty years

In the first item, the proficient student may be able to reject distractors A and C, so he has a 50-50 chance of getting this item correct. On the second item, he is able to reject all three distractors, so he is expected to get this item correct. Over the two items the group of minimally proficient students would be expected to have a mean score of 1.5. Out of two items this is a proportion of .75 which would be the estimate of the proficiency standard. Of course an accurate and valid estimate of the standard depends on the number of items from the domain that are rated and the amount of agreement among the raters.

Another method is dependent on sound test construction procedures wherein items are targeted at the minimally proficient population. For instance, if one has remedial

English courses, the items should be so designed that a person passing those courses would have a 50-50 chance of getting each item correct. It is important that the 50% level be chosen so that maximum discrimination will occur at the most crucial point along the ability scale. Based on the Nedelsky method, this 50% level would correspond to designing distractors so that the minimally proficient student would be able to reject all except two. If one is sure that the ability of the test population is normally distributed, this 50% criteria would be reflected in sample item difficulties close to .5, but if the distribution is skewed as in the UCLA data, then item difficulty would be a negatively biased indicator of the efficiency of an item.

Although multiple-choice items have been used to present these methods, they can be applied to other types of test items such as fill-in-the-blank, dictation, or even FSI interviews; just as long as the test designer can specify how the minimally proficient examinee is expected to perform on each item.

Determining the Best Cut-Score

Once a proficiency standard has been established, a cut-score must be derived. Because test scores are fallible, no matter what cut-score is chosen, errors of classification will occur. In some cases, a proficient examinee will obtain a score lower than the cut-score and

therefore fail the exam, while in other cases non-proficient students will score higher than the cut-score simply due to luck. The problem now becomes one of finding a cut-score which will keep these errors below acceptable risk levels. Stig Fhaner (1974) suggests a method which makes use of the binomial distribution. This distribution describes the chances of a person with a given ability of obtaining a particular score. The assumptions of the model are that the items be scored as either right or wrong (1's and 0's), that the items be independent of each other, and that they measure the same thing. In order to use the formula, we need to specify three things, the number of items on the test, the proficiency level of the examinee and the observed score. The model would be expressed as in Equation 1.

$$(1) \quad p(x) = \binom{n}{x} a^x (1-a)^{n-x}$$

where: p(x)=probability of obtaining score x
n=number of test items
x=observed score
a=ability or proficiency level, $0<a<1$

For example if we had a 10 item test and the examinee has an ability of answering 3/4 of the items in the test domain, the probability of getting a score of 8 is set up as in Equation 2.

$$(2)\quad p(8) = \binom{10}{8} .75^{8} (1-.75)^{10-8} = .28$$

Similarly, the probability of scoring a 9 is .19 and a 10 is .06. The probability of scoring an 8 or higher is then .28+.19+.06=.53. Using this formula, Fhaner describes the best cut-score as that which will pass no more than some small percentage, say 1%, of the persons who are definitely non-masters and fail no more than a certain percent of masters. We can then set up Equations 3 and 4 and solve for c.

$$(3)\quad p(x \geq c \mid a_1) = \sum_{x=c}^{n} \binom{n}{x} a_1^{x} (1-a_1)^{n-x} < \text{alpha}$$

$$(4)\quad p(x < c \mid a_2) = \sum_{x=0}^{c-1} \binom{n}{x} a_2^{x} (1-a)^{n-x} < \text{beta}$$

where: $p(x \geq c \mid a_i)$ = probability of getting a score greater than or equal to c given ability a_i.
a_1 = ability definitely insufficient for passing
a_2 = ability definitely sufficient for passing
alpha = tolerable risk of passing a non-master
beta = tolerable risk of failing a master

Equation 3 specifies that the probability of getting a score higher than the cut score, when one's true ability is less than the minimum proficiency standard, is less than or

equal to the percentage risk, alpha, while Equation 4 expresses the probability of getting a score less than the cut-score, when one's true ability is higher than the standard, is less than or equal to the risk, beta.

Since the binomial distribution can be approximated by the normal distribution when the number of test items is large and the ability parameter, $\underline{a}$, is not near 0 or 1 we can set up Equations 5 and 6.

$$(5) \quad Z_{beta} = \frac{c + .5 - n(a_2)}{\sqrt{n(a_2)(1-a_2)}}$$

$$(6) \quad Z_{alpha} = \frac{c - .5 - n(a_1)}{\sqrt{n(a_1)(1-a_1)}}$$

where: Z_{beta} = Z-score normal deviation corresponding to the beta risk

Z_{alpha} = Z-score normal deviation corresponding to the alpha risk

On the UCLA 88 item multiple-choice test one might choose .49 to be the ability definitely insufficient for passing and .7 to be the ability definitely sufficient for passing with a 1% risk of passing a non-master and a 10% risk of failing a master. The equations would be set up as in 7 and 8.

In these equations a Z-score of 2.33 corresponds to an alpha of .01 and a Z-score of -1.28 corresponds to a beta of .1. The cut-score which satisfies the proficiency standards and the risks of misplacement is calculated to be 55.

$$(7)\quad -1.28 = \frac{c + .5 - 88(.70)}{\sqrt{88(.70)(.30)}} \rightarrow c \doteq 55$$

$$(8)\quad 2.33 = \frac{c - .5 - 88(.49)}{\sqrt{88(.49)(.51)}} \rightarrow c \doteq 55$$

External Criterion Checks

It should be noted that these cut-scores were determined without the use of any external criterion or in reference to the distribution of sample scores. If one were to have complete confidence in predicting the test performance of particular ability groups, then there is no need for any other criteria. However, this is rarely the case and it is helpful to check one's classifications against examinee performance in subsequent related activities.

Criteria as specific as performance on a complex language task such as a cloze test or a dictation may be used, or something more general may be selected, such as grade point average. Using the proposed cut-score of 55 as the basis for classifying persons into proficient and non-proficient categories, it is possible to determine the corresponding scores on the 1977 cloze test and dictation tests that minimize the number of misclassifications. This is an idea suggested by Rand Wilcox (1977).

For example, if we choose a cloze score of 35 as a standard, we can form the 2 x 2 cross-classification that appears in Table 3.

INSERT TABLE 3 ABOUT HERE

In the upper left cell are the number of persons classified as non-masters by both the multiple-choice and cloze test and in the lower right cell are the number of persons classified as masters by both tests. In the opposite diagonal cells are the number of misclassifications. Here, 2 persons were classfied as masters by the cloze, but non-masters by the multiple-choice test, while 248 persons were masters by the multiple-choice criterion but non-masters by the cloze. It is these misclassifications that are to be minimized.

In Figure 7 are graphed the number of misclassifications as a function of the external criterion scores. On the cloze test the minimum occurs when a score of 22 is used as the cut point. This roughly corresponds to 44% of the cloze items, which is a reasonable expected performance of a proficient examinee (see, for example, Bormuth, 1967, p.18). Figure 8 shows that on dictation the minimum number of misclassifications occurs at a score of 30, indicating that the proficient examinee would be expected to get about 3 out of 5 words, which, although not great, is still over half.

It is also possible to use external criteria on subsamples of the test scores. For these data, grade point

INSERT FIGURES 7 AND 8 HERE

averages were availible for 99 graduate engineering students from their first two years of graduate study, and could be used as an external criterion. Since 3.25 is considered to be the passing grade point average for a graduate student at UCLA, this was used as a fixed criterion for declaring mastery and non-mastery status, and different scores on the multiple-choice test were proposed in an attempt to minimize classification errors.

Figure 9 indicates the best cut-score is 43, which is much lower than the proposed cut-score of 55. Two interpretations are possible: either the proficiency standard on the multiple-choice test is too high for indicating the ability to do adequate work as an engineering graduate, or the remedial English courses that these students received upon failing the exam helped them to pass their regular university courses. In any case 43 out of 88 items does just about correspond to the .5 ability level which was originally put forth as the minimum level of proficient performance.

INSERT FIGURE 9 ABOUT HERE

Conclusion

To sum up, norm-referenced cut-score setting is a method which can be used when populations and test forms are stable. It should not be used when one is unsure of the population distribution of ability or when one suspects test form by person interactions. It is not a substitute for having to make judgements about how test performance relates to the proficiency level of interest.

In effect, then, there are four areas of concern in setting a cut-score: 1) making sure that the test covers the abilities of the examinees, 2) determining the proficiency standard, 3) acknowledging that there will be errors of measurement and that risks of making misclassifications be specified, and 4) using external criteria for empirical verification of the proficiency standard. No matter what type of testing format is used, these four concerns must be addressed in arriving at a defensible cut-score.

References

Bormuth, J. Cloze readability procedure. Center for the Study of Evaluation. CSE Report #4, University of California, Los Angeles, 1967.

Phaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 24, 172-75.

Wedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19. 1954, 14, 3-19.

Wilcox, R. Determining the optimal passing score in the two-way classification problem. Unpublished manuscript. Center for the Study of Evaluation, University of California, Los Angeles, 1977.

Table 1

1977 Subtest Reliabilities and Intercorrelations

	Dictation	Cloze	Multiple-Choice
Dictation	.94		
Cloze	.73	.88	
Multiple-Choice	.85	.77	.95

Table 2

1978 Subtest Reliabilities and Intercorrelations

	Dictation	Cloze	Multiple-Choice
Dictation	.97		
Cloze	.56	.97	
Multiple-Choice	.54	.79	.95

Table 3

1977 2x2 Cloze by Multiple-Choice
Cross-Classification of Mastery Status When Cloze
Cut-Score is 35 and Multiple-Choice Cut-Score is 55

		Cloze		
		Non-mastery	Mastery	
Multiple Choice	Non-mastery	174	2	176
	Mastery	248	36	284
		422	38	460

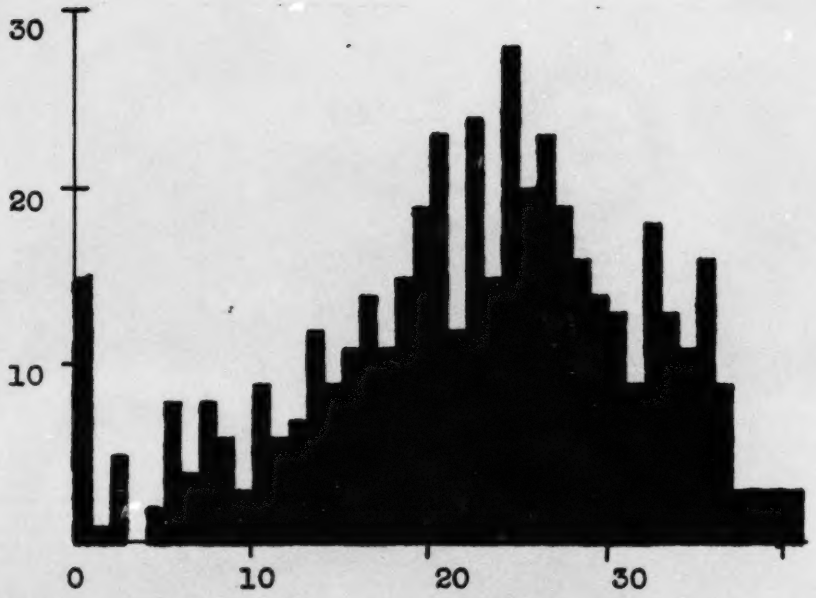

Fig. 1. 1977 distribution of scores on 50-item cloze, N=460, $\bar{X}$=21.88, s=9.38.

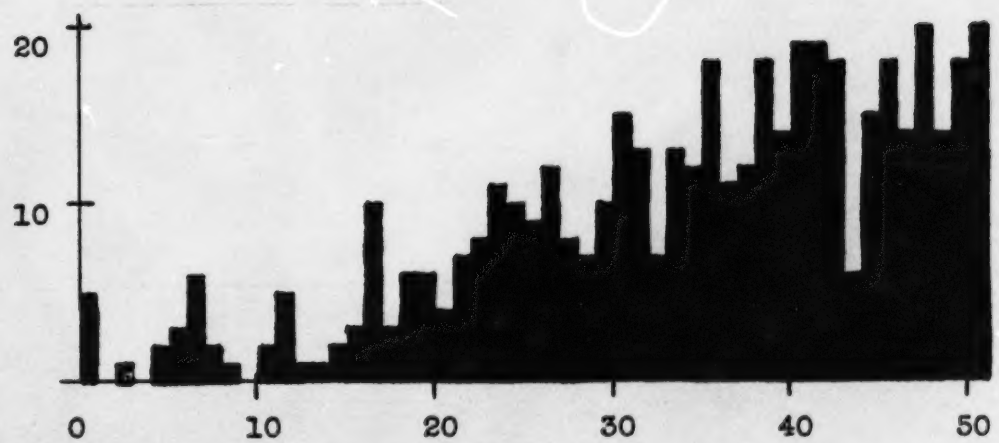

Fig. 2. 1977 distribution of scores on 50-word dictation, N=460, X=33.99, s=11.90.

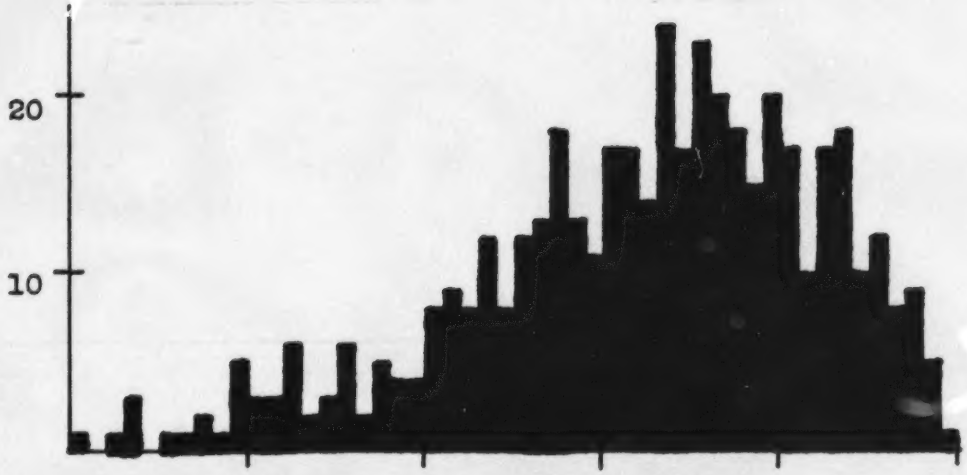

Fig. 3. 1977 distribution of scores on 88-item multiple-choice test, n=460, $\bar{X}$=57.06, s=15.92.

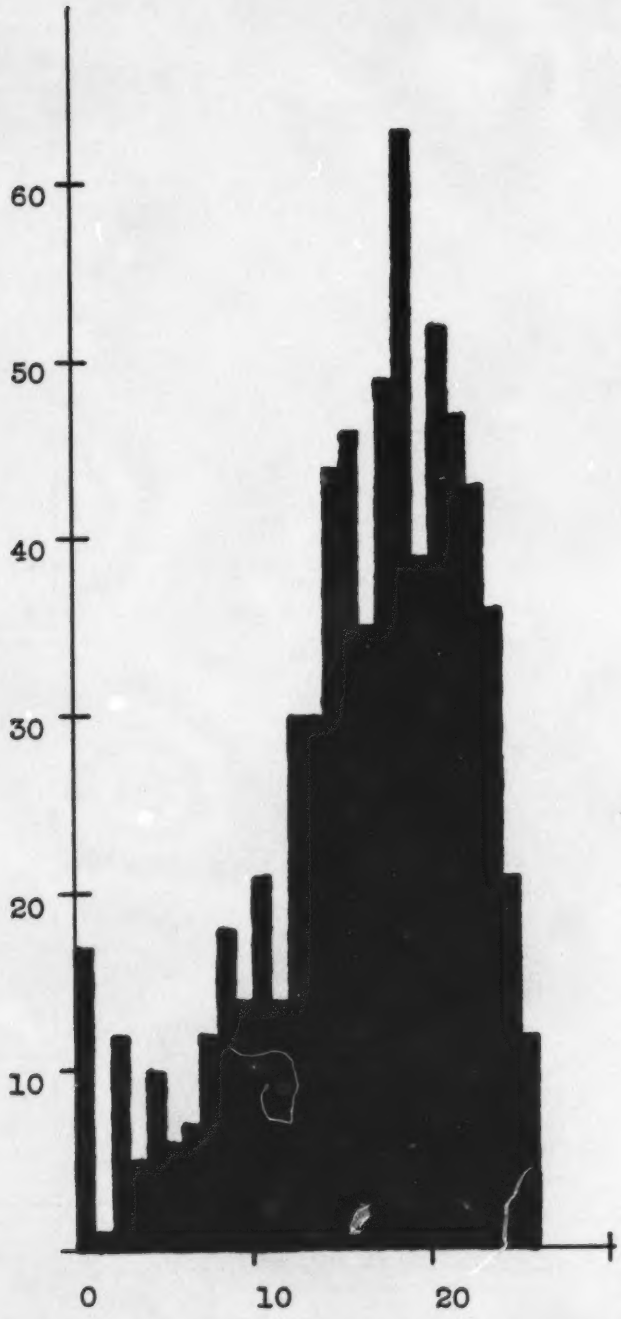

Fig. 4. 1978 distribution of scores on 25-item cloze, N=689, $\bar{X}$=16.39, s=9.19.

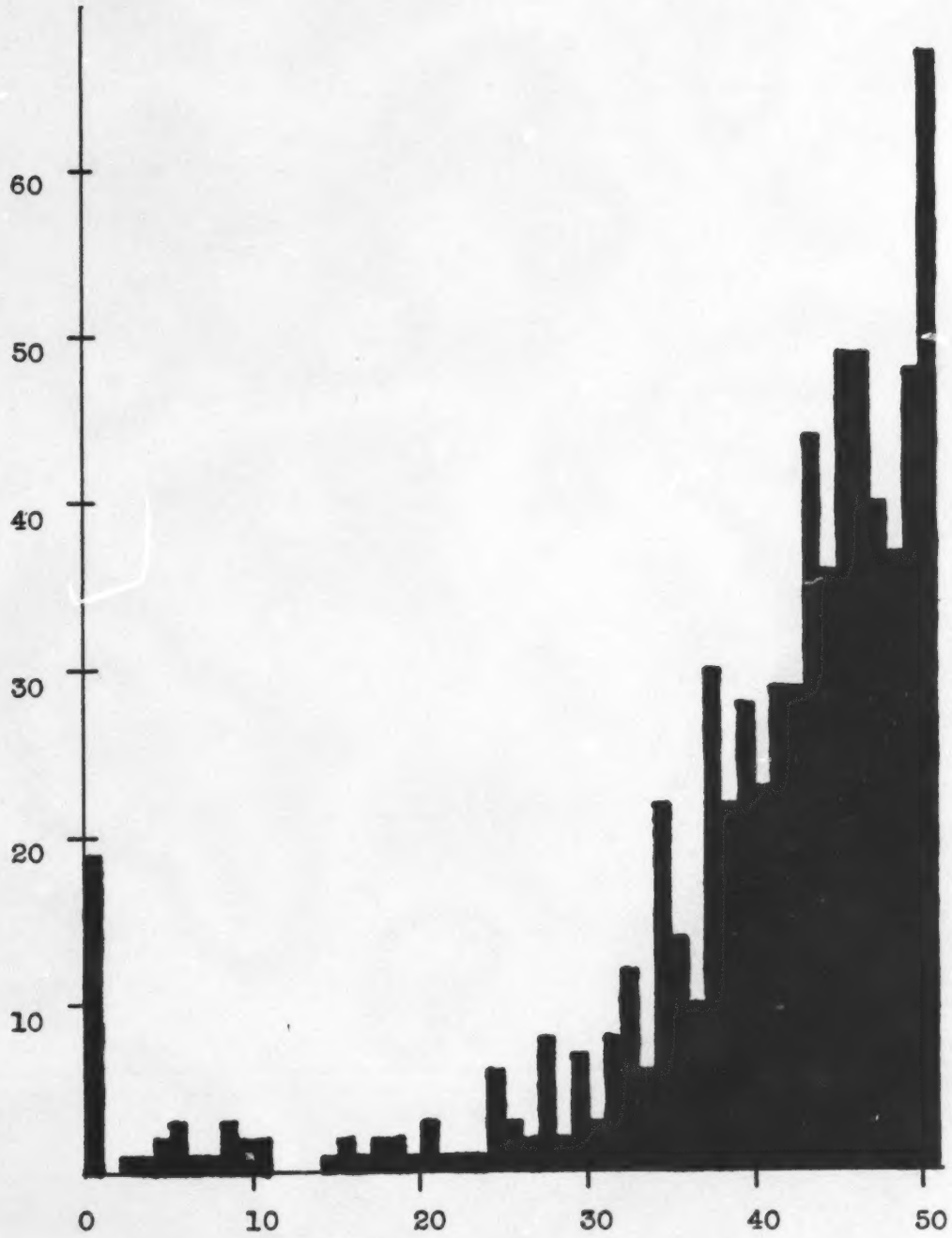

Fig. 5. 1978 distribution of scores on 50-word dictation, N=689, $\bar{X}$=40.34, s=12.10.

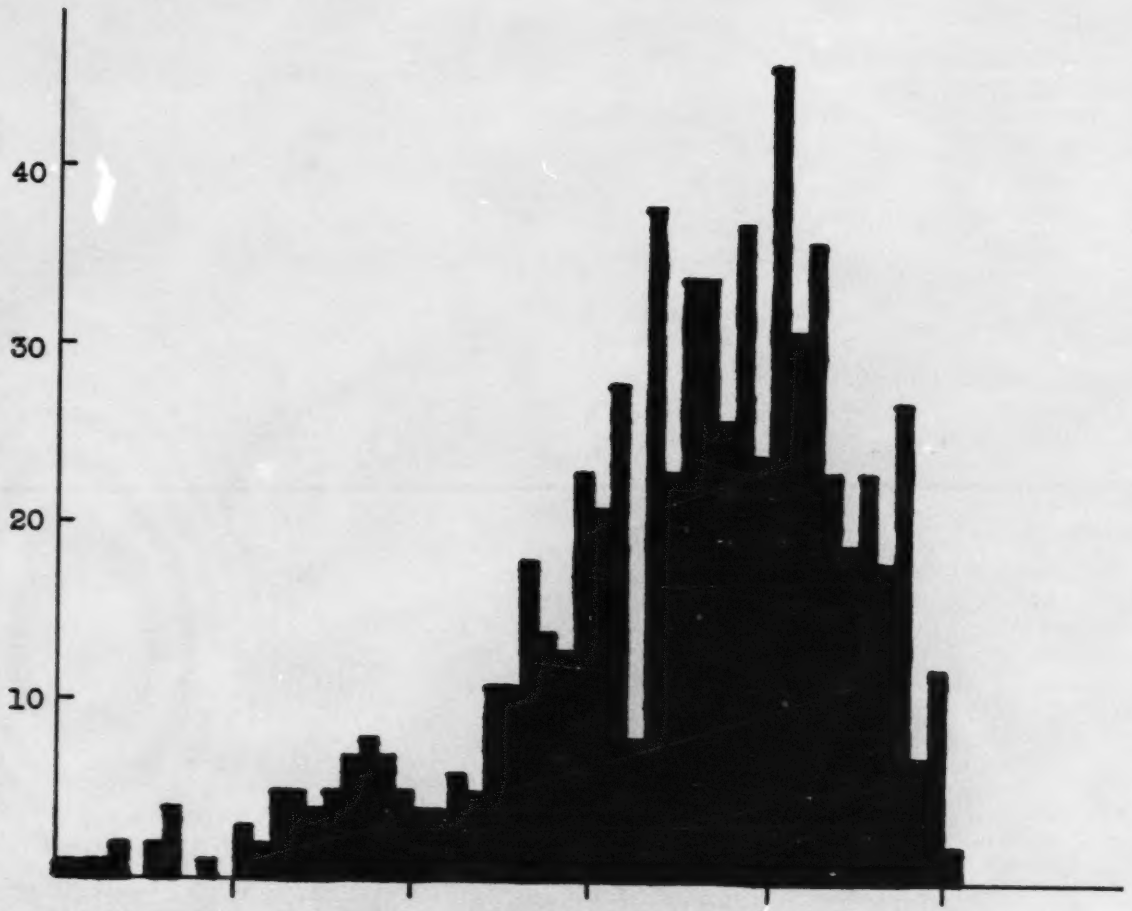

Fig. 6. 1978 distribution of Z-scores on 135-item multiple-choice test, N=689, $\bar{X}$=89.59, s=23.22.

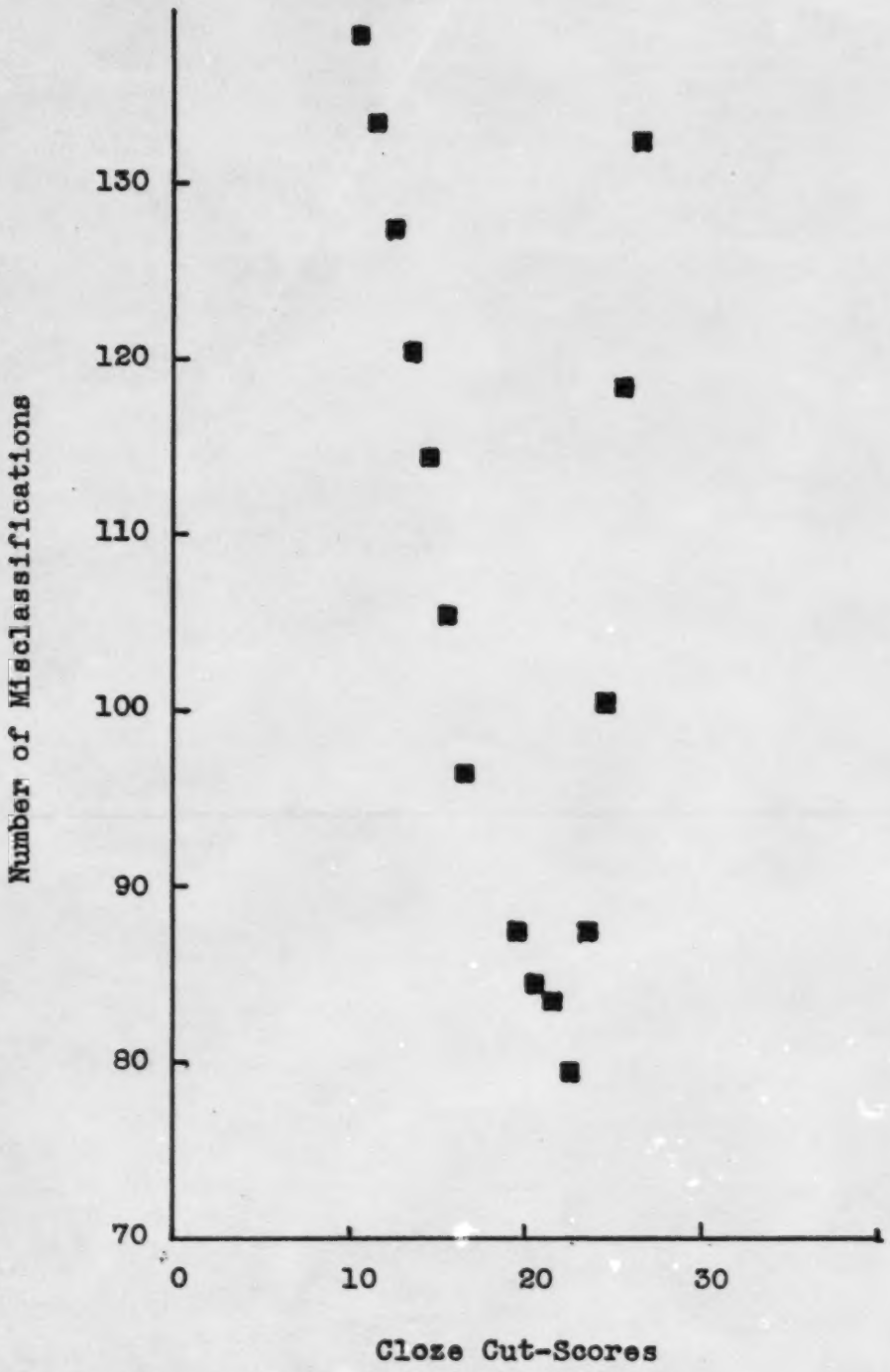

Fig. 7. Number of misclassifications in 2x2 cloze by multiple-choice cross-classifications of mastery status when multiple-choice cut-score=55 and cloze cut-scores vary.

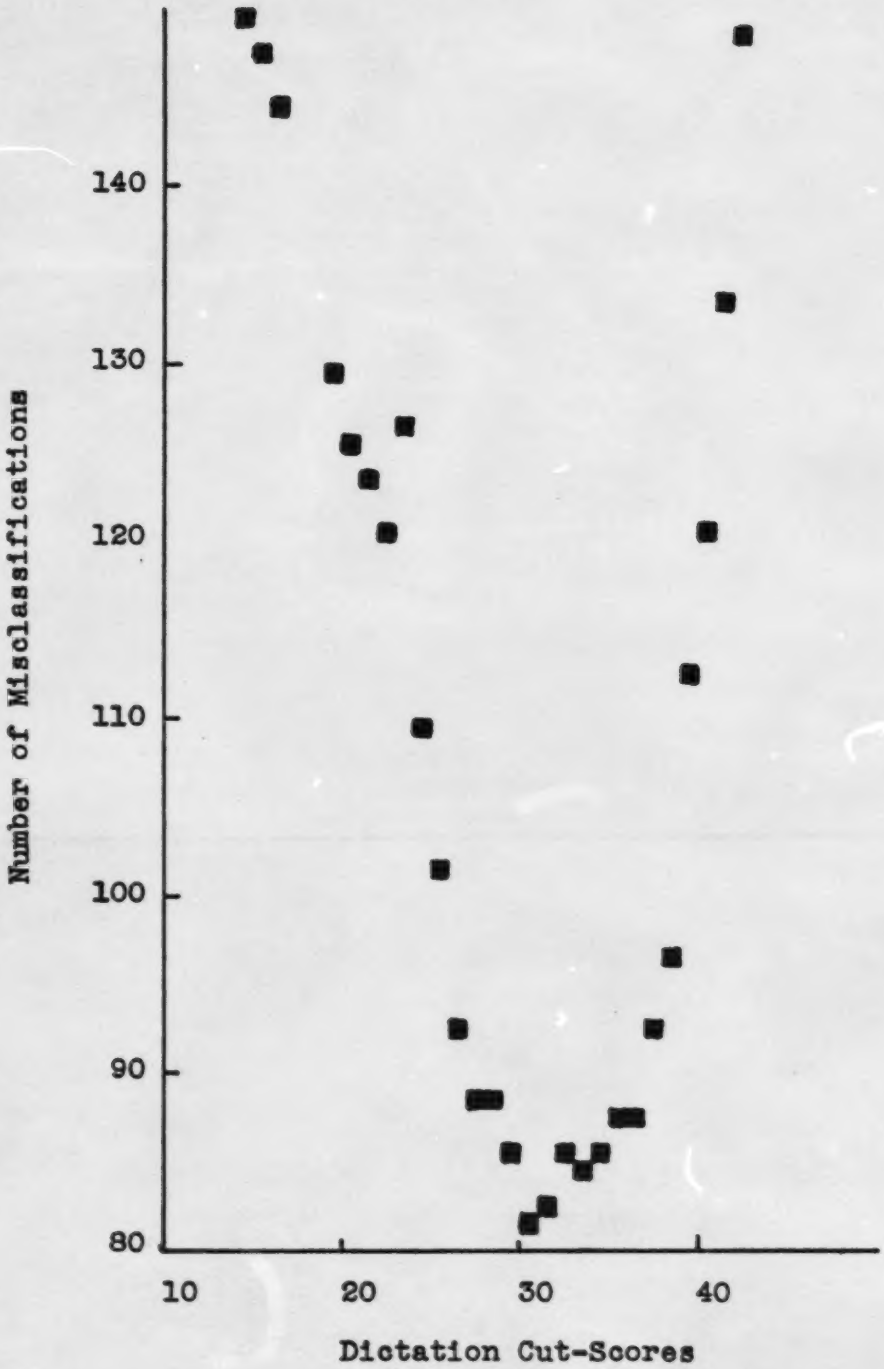

Fig. 8. Number of misclassifications in 2x2 dictation by multiple-choice cross-classifications of mastery status when multiple-choice cut-score=55 and dictation cut-scores vary.

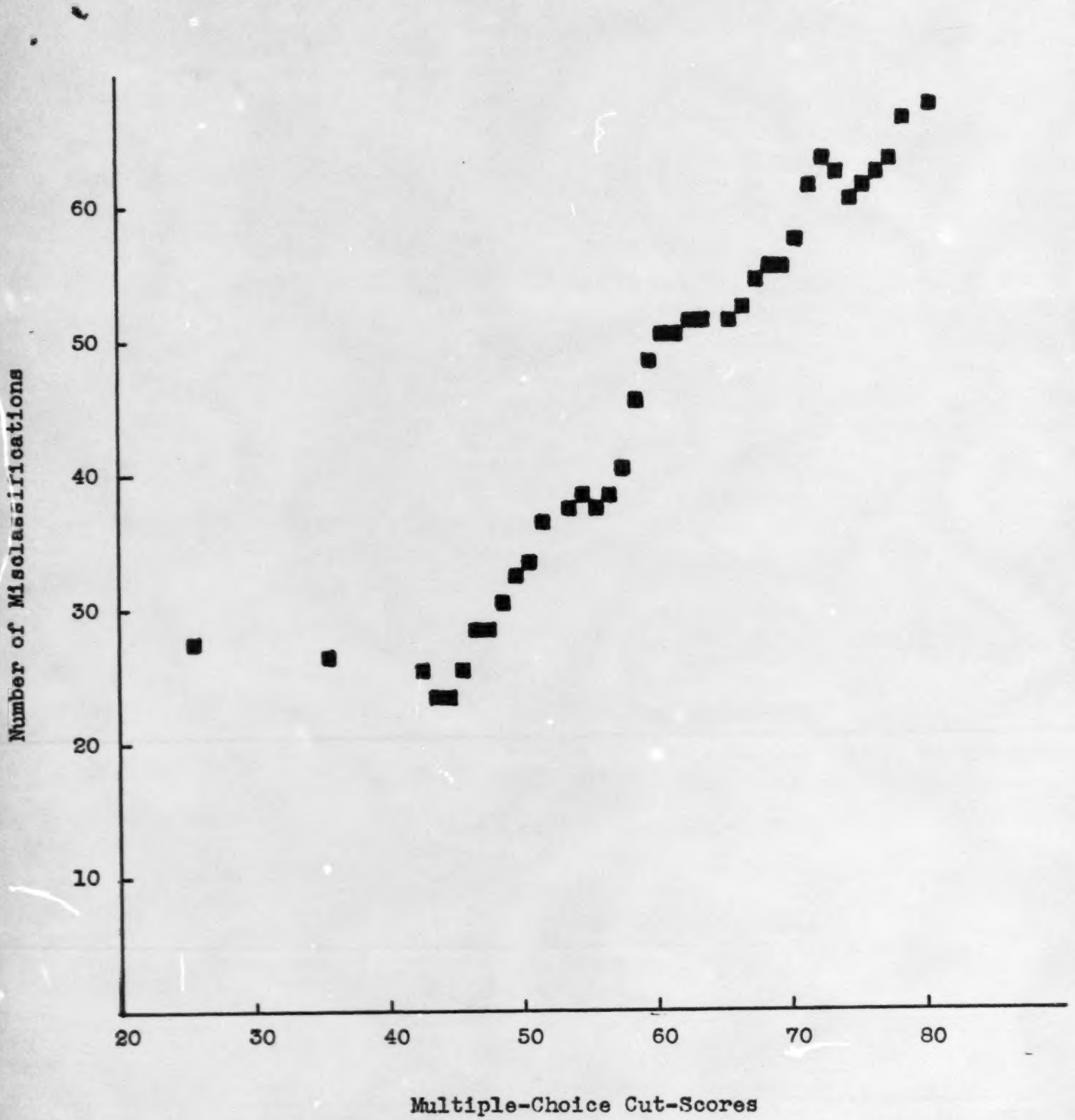

Fig. 9. Number of misclassifications in 2x2 GPA by multiple-choice cross-classifications of mastery status among 99 engineering graduate students when GPA criterion=3.25 and multiple-choice cut-scores vary.